ЭЛЕКТРОННЫЕ ИНФОРМАЦИОННЫЕ РЕСУРСЫ

УДК 347.77:025.4.03 + 81'322.2
https://doi.org/10.33186/1027-3689-2025-5-58-80

Кластерный подход к формированию наборов патентных данных и оценивание качества поиска «уровня техники»

А. В. Горбунов

Федеральный институт промышленной собственности,
Москва, Российская Федерация,
gorbunov@rupto.ru

Аннотация. По мере расширения мирового патентного фонда возрастает и сложность поиска уже опубликованных патентных документов для оценки новизны технических решений – так называемого извлечения «релевантного уровня техники», «предшествующего уровня техники» или «уровня техники» из общедоступных патентных данных. Поиск такой информации связан со значительными трудностями, обусловленными её объёмом и сложностью. Результаты ряда исследований свидетельствуют о растущем масштабе использования машинной обработки естественного языка (NLP) для повышения точности и комплексности патентного поиска. Несмотря на достигнутые успехи, до сих пор не представлено системы автоматического патентного поиска, способной демонстрировать приемлемые точность и полноту. Автор статьи считает, что развитие новых, эффективных подходов к построению таких систем существенно ограничивается недостатком подготовленных наборов данных для обучения и тестирования. Автоматизированное создание наборов данных произвольной конфигурации – с учётом различных критериев отбора (документы одного или нескольких патентных ведомств; все опубликованные документы за ограниченный период времени; виды документов; классы патентной классификации и т. д.) – позволит снять ограничения и создавать наборы данных, соответствующие потребностям и целям разработчиков систем автоматического патентного поиска. В статье предложены новые подходы как к созданию наборов данных для обучения и тестирования систем автоматического патентного поиска уровня техники, так и к оценке эффективности созданных систем.

Ключевые слова: патентный поиск, поиск по известному уровню техники, наборы данных, патентная коллекция

Благодарности: автор выражает благодарность доктору педагогических наук, профессору Н. В. Лопатиной, доктору экономических наук О. П. Неретину, кандидату технических наук Б. Л. Генину за неоценимую методологическую помощь в подготовке статьи; Д. С. Золкину, С. В. Сапожникову, И. В. Некрасову за разработку программ и подготовку наборов данных.

Для цитирования: Горбунов А. В. Кластерный подход к формированию наборов патентных данных и оценивание качества поиска уровня техники // Научные и технические библиотеки. 2025. № 5. С. 58–80. https://doi.org/10.33186/1027-3689-2025-5-58-80

DIGITAL INFORMATION RESOURCES

UDC 347.77:025.4.03 + 81'322.2
https://doi.org/10.33186/1027-3689-2025-5-58-80

The cluster approach to acquiring patent datasets and assessing the quality of "prior art search"

Alexander V. Gorbunov

Federal Institute of Industrial Property, Moscow, Russian Federation, gorbunov@rupto.ru

Abstract. As the global patent collection is widening, the complexity of patent documents search for assessing technique novelty, i. e. revealing the relevant art or prior art from public patent data, is increasing, too. Searching for this information, vast and complex, is challenging. Research findings evidence on the increasing scale of NLP use for more accurate and integrated patent search. Despite many achievements, the automated patent search system for appropriate accuracy and completeness has not been introduced. The author argues that development of new effective approaches to designing these systems is significantly limited due to the lack of the datasets ready for educating and testing. The automated acquisition of datasets of arbitrary configuration (with consideration for various selection criteria, i. e. documents by patent agency/agencies; all published documents for a limited period of time: document types; patent classification classes,

etc.) would enable to eliminate limitations and build the datasets meeting the needs and goals set up by the systems designers. The author proposes new approaches to dataset acquisition, testing of automated art patent search systems, and assessment of these systems.

Keywords: patent search, prior art search, dataset. patent collection

Acknowledgements: The author extends his thanks to Prof. N. V. Lopatina, Dr. Sc in Pedagogy, O. P. Neretin, Dr. Sc. in Economics, B. L. Genin, Cand. Sc. in Engineering, for their invaluable methodological support of the publication; D. S. Zolkin, S. V. Sapozhnikov and I. V. Nekrasov for developing the software and the datasets.

Cite: Gorbunov A. V. The cluster approach to acquiring patent datasets and assessing the quality of "prior art search" // Scientific and technical libraries. 2025. No. 5, pp. 58–80. https://doi.org/10.33186/1027-3689-2025-5-58-80

1. Введение

Патент – это исключительное право на производство, использование или продажу изобретения, выдаваемое государственными патентными ведомствами [**1**]. Чтобы получить патент, заявитель должен доказать, что его (её) идея обладает новизной, техническим уровнем и промышленной применимостью. Обычно основанием для такого утверждения служит сравнение заявки с уже известными техническими решениями, описания которых публикуются в базах данных патентных ведомств и других специализированных организаций. По мере экспоненциального развития технологий происходит увеличение числа патентных заявок и выданных патентов, что приводит к разрастанию баз патентной информации [**2**, **3**].

Патентный поиск (PR) – это направление информационного поиска (IR), включающее создание стратегий и подходов, позволяющих идентифицировать релевантные патентные документы в ответ на заданный поисковый запрос [**4**]. Патентный поиск применяется для решения задач сферы интеллектуальной собственности, связанных с выявлением охраняемых технических решений, которые отражают актуальные технические достижения (так называемый «уровень техники») в рассматриваемой отрасли в соответствии с запросом.

Традиционные методы ручного патентного поиска, хотя и хорошо проработаны, являются трудоёмкими и отнимают много времени, что зачастую приводит к неполным и противоречивым результатам. Даже хорошо подготовленные и опытные патентные поверенные считают, что выявление предшествующего уровня техники является утомительной и трудозатратной задачей [**5**], при этом существует значительный риск пропустить важные документы [**6**, **7**].

Автоматизация патентного поиска эволюционировала от поиска по ключевым словам к сложному контекстному поиску, основанному на глубоком обучении (DL) и обработке естественного языка (NLP).

Технологии NLP позволяют распознавать и анализировать семантические связи в патентных текстах, выдавая результаты поиска, которые являются релевантными и соответствуют контексту. Аналогичным образом глубокое обучение помогает автоматизировать анализ последовательностей и извлечение признаков, повышая скорость и масштабируемость поисковых систем. Эти передовые методы значительно превосходят традиционные подходы по эффективности, точности и масштабируемости. Благодаря лучшему пониманию семантики патентных текстов они не только повышают точность результатов, но и упрощают процесс поиска.

Совершенствование технологий и, как следствие, быстрое изменение патентной лексики, требуют всё больше новых и постоянно обновляемых наборов данных для обучения и тестирования поисковых систем, основанных на NLP и DL. До настоящего времени наиболее используемыми и признаваемыми сообществом являются тестовые коллекции документов Межъязыкового форума по оценке интеллектуальной собственности (CLEF-IP) [**8–11**], исследовательского проекта Японского национального института испытательных полигонов информатики и сообщества по доступу к информации (NTCIR) [**12–15**] и конференций по оценке систем текстового поиска (TREC) Американского института стандартов (NIST) [**16–18**]. Все три коллекции создавались и обновлялись в течение нескольких лет, но в настоящее время не сопровождаются. Тщательный отбор документов для каждой коллекции с широким привлечением экспертов, безусловно, позволил обеспечить качество, но в текущих условиях ситуация осложняется всё расширяющимся охватом отраслей. Нужно отметить, что и в случае упомянутых коллекций отбор данных «вручную» тоже явился ограни-

чивающим фактором – как с точки зрения отраслевого охвата (например, TREC-CHEM – содержит только патенты по химии), так и с точки зрения языков (NTCIR – японский и английский язык, остальные – только английский язык).

Мировая коллекция патентной информации (WPI[1]), представленная Lupu, Bampoulidis и Papariello [**19**], охватывает опубликованные за два года (2014 и 2015) патенты всех основных органов: Европейского патентного ведомства (код ведомства – EP), ведомства США по патентам и товарным знакам (US), Всемирной организации интеллектуальной собственности (WO), Китайского патентного ведомства (CN), Японского патентного ведомства (JP) и Корейского патентного ведомства (KR). Коллекция содержит полный набор библиографических данных, полные тексты всех документов, а также все необходимые изображения и дополнительные материалы. Коллекция, подготовленная во многом «вручную», является статичной и не пополняется. Кроме того, она не содержит тестового набора данных или разметки документов. Пользователю необходимо определять тестовый набор для машинного обучения самостоятельно.

Таким образом, до настоящего времени специалистам по машинному обучению не было предложено подхода и инструментов для построения наборов патентных данных произвольной конфигурации в аспекте временного охвата, языков, отраслей и т. д. В этой статье описан подход к автоматизированному созданию таких наборов, позволяющий избегать «бутылочного горлышка» в отборе документов вручную, раскрываются авторская методика и показатели качества патентного поиска, расширяющие набор инструментов оценки точности и полноты в соответствии с задачами патентного поиска.

2. Материалы и методы

Источниковой базой исследования выступал Государственный патентный фонд[2] (ГПФ) как «часть государственного ресурса научно-технической информации, предназначенная для удовлетворения потребностей в патентной информации всех категорий пользователей…

---

[1] Размещена на Zenodo как защищённая коллекция, доступна по следующей ссылке: https://doi.org/10.5281/zenodo.1489994.

[2] https://rospatent.gov.ru/ru/state-patent-collection

представляющая собой совокупность систематизированных и снабжённых справочно-поисковым аппаратом источников информации, относящихся к изобретениям, полезным моделям, промышленным образцам, товарным знакам, наименованиям мест происхождения товаров, географическим указаниям, программам для ЭВМ, базам данных и топологиям интегральных микросхем и включающих патентную документацию и непатентную литературу» [**20**].

Во многих случаях в ГПФ патенты промышленно развитых стран опубликованы, начиная с патента № 1. Большая часть документов (более 135 млн) доступны в электронном виде и могут быть извлечены через API Поисковой платформы Роспатента в формате XML.

Для формирования наборов патентных данных – как обучающих, так и тестовых – в этом исследовании используется особенность структуры опубликованных патентов [**21**] – наличие поля ИНИД (56) «Документы уровня техники», то есть указание экспертов на те документы, которые были оценены и учтены как документы уровня техники при принятии решения о выдаче патента. Несмотря на то, что публикуются не все заявки и не все патенты, а также публикуется много документов, связанных с выдаваемым в итоге патентом, но не содержащих полной информации о патентуемом решении, можно утверждать, что не менее 25% мирового патентного фонда составляют документы, содержащие поле (56), то есть размеченные экспертизой, а это около 30 млн документов. Такой объём достаточен для подготовки любых наборов данных – как для обучения, так и для тестирования поисковых систем.

В связи с этим интересна публикация международного исследовательского проекта IRF (Information Retrieval Facility) «Задача поиска кандидатов» (Prior Art Candidates Search Task, PAC) [**22**]. Проект IRF опирался на материалы конференции CLEF-IP. Среди прочего в работе было предложено использовать для оценки методов патентного поиска информацию о цитировании, имеющуюся в патентных документах. При этом имелись в виду как цитаты заявителя, так и цитирование экспертизы. По мнению Giovanna Roda, Veronika Zenz, Mihai Lupu, использование цитат патентных экспертов ведомства в оценке релевантности документов выглядит более предпочтительным, чем привлечение сторонних экспертов для «ручной» оценки релевантности [**23**].

Таким образом, цитаты из поля (56) уже признаны сообществом в качестве разметки. В настоящей работе сделан следующий шаг – в ка-

честве релевантных документов признаются и документы, входящие в так называемые «простые патентные семейства».

Использование поля (56) в качестве «источника правильных ответов» для обучения систем поиска или тестирования таких систем представляется удачной альтернативой «ручному» поиску либо оцениванию релевантности документов уровня техники. Однако при проверке совпадений необходимо учитывать, что ссылка, указанная экспертом в поле (56), характеризует не только и не столько конкретный патентный документ, но в большей мере изобретение, на которое подана заявка или уже выдан патент. Действительно, эксперт ведомства может процитировать в поле (56) не все найденные релевантные документы, например, в самом простом случае – патентным экспертом может быть выявлена пара релевантных документов (заявка и патент), характеризующих одно и то же изобретение. В поле (56) эксперт укажет ссылку только на один из этих документов. В связи с этим достаточно, чтобы указанная ссылка на патентный документ, описывающий некоторое изобретение, и одна из ссылок, найденных системой, относились к одному и тому же изобретению. Иными словами, при проверке совпадения нужно сравнивать не просто идентификаторы конкретных документов, но и совпадения любых идентификаторов патентных документов, входящих в семейство патентов – аналогов заявки – образца, и, соответственно, любых идентификаторов патентных документов, входящих в семейство патентов – аналогов найденных документов.

Важность учёта патентных документов, входящих в семейство патентов – аналогов заявки, отмечает и Maik Fröbe и др. [**24**].

Формирование наборов данных для обучения и тестирования систем автоматизированного поиска уровня техники

Сказанное выше послужило основой кластерного подхода к созданию наборов данных для обучения и тестирования автоматизированных систем патентного поиска.

Цель подхода – создать такой набор данных, чтобы для каждого документа (назовём такие документы базовыми) было собрано подмножество документов, по мнению экспертизы, определяющих уровень техники. Для этого в набор данных в качестве базовых выбираются те документы коллекции, которые содержат поле (56), то есть патенты, а также заявки (если таковые были опубликованы), на основании кото-

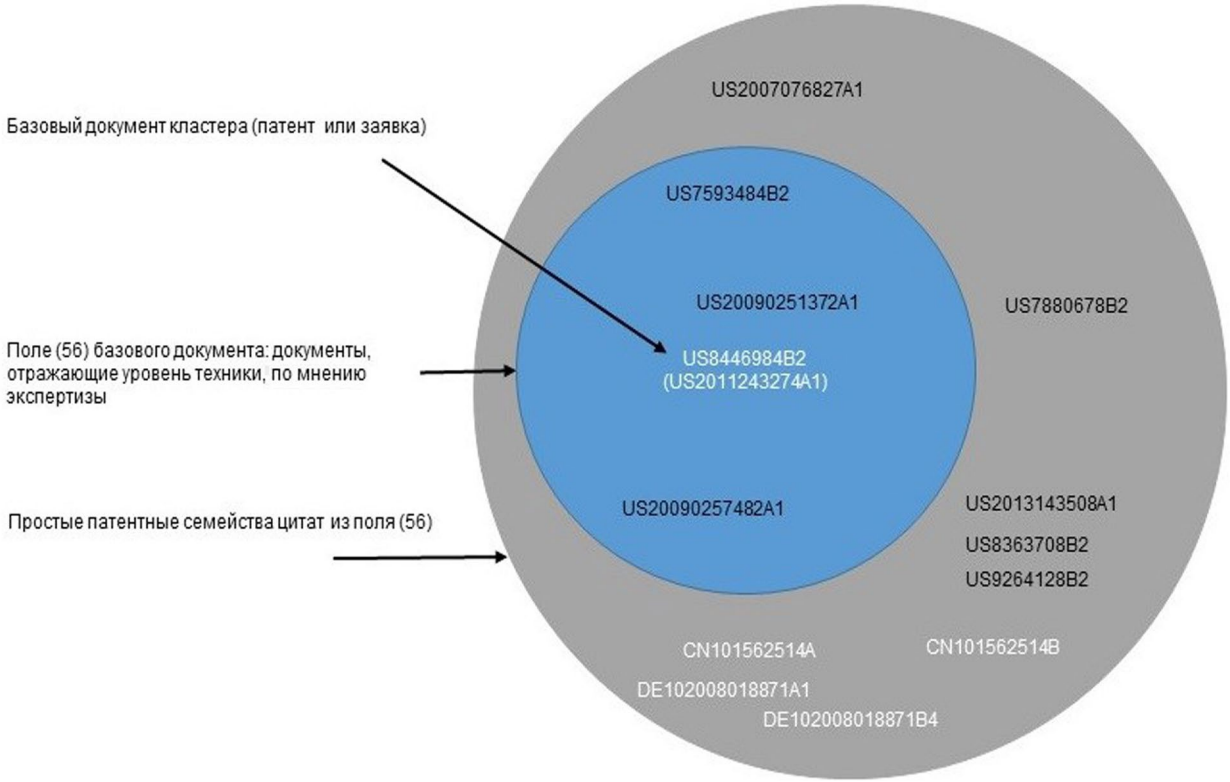
рых были выданы эти патенты. Набор данных дополняется документами, которые указала экспертиза в поле (56), а также документами, входящими в их патентные семейства. Таким образом вокруг каждого базового документа в наборе формируется кластер – всё множество документов, которые описывают изобретения, отражающие, по мнению экспертизы, уровень техники для базового документа. Более строгое определение кластера патентных документов приведено в работе автора и коллег [**25**].

Кластерный подход к формированию наборов данных позволяет практически полностью автоматизировать процесс их создания, исключить «ручную» составляющую по подбору документов уровня техники. В то же время в результирующем наборе данных будут гарантированно собраны все документы, которые эксперты патентного ведомства посчитали (или могли посчитать) документами уровня техники.

В таком подходе формулировка задачи поиска уровня техники видоизменяется – от «в наборе данных найти документы, отражающие уровень техники для рассматриваемого документа» к «найти кластер документов, наиболее подходящий для рассматриваемого документа».

В случае обучения и тестирования, когда наборы данных уже определены и подготовлены, из них извлекаются базовые документы и заявки к ним. Затем по какой-либо из извлечённых заявок или по массиву таких заявок проводится поиск. Пример кластера, построенного для патента US8446984B2, представлен на рисунке.

В кластер вошло 13 документов: заявка, по которой был выдан этот патент, документы, указанные в поле (56) патента и их патентные семейства (показаны во внешней окружности). При обучении системы поиску, например, по документам, выданным американским патентным ведомством, документы кластера, выданные другими ведомствами, оказавшиеся в ранжировании системы, не будут учтены. Такие документы на рисунке выделены белым цветом. Таким же образом выделены сам патент и его аналог (в данном случае заявка, по которой он выдан) – нахождение этих документов тоже не будет учтено при оценке качества поиска.



Кластер патента US844698B2. Оценка качества поиска

Информационный поиск «в целом» и патентный поиск как разновидность информационного поиска направлены на максимизацию извлечения из коллекции релевантных документов (истинно положительные, TP), минимизацию игнорирования релевантных документов (ложноотрицательные, FN) и минимизацию извлечения нерелевантных документов, ошибочно идентифицированных как релевантные (ложноположительные, FP). Показатели recall, precision, F1-score, Mean Average Precision (MAP), accuracy и другие обычно используются в литературе для оценки эффективности систем информационного поиска [**26–28**]. Все они опираются на соотношения TP, FN, FP и TN (истинно отрицательные), при этом подразумевается, что релевантные документы известны заранее (то есть к моменту вычисления метрики), а набор релевантных документов является исчерпывающим – во всяком случае, для той коллекции документов, в которой осуществляется поиск.

Для наборов данных патентных документов, сформированных вручную, релевантные патенты, оценённые и отобранные экспертами, становятся целью для системы поиска уровня техники.

Если набор данных формируется автоматически, исходя из того, что релевантными являются документы, указанные в поле (56) и документы, вошедшие в соответствующий кластер, то любой найденный при автоматическом поиске уровня техники патентный документ можно считать релевантным базовому документу, если он входит в семантический кластер, в состав которого входит этот базовый документ.

Необходимо учитывать, что документы кластера описывают изобретения и, хотя эксперт указал в поле (56) всего один документ для каждого изобретения в кластере, вполне вероятно, что патентное семейство любого из этих, указанных в поле (56) изобретений будет шире, то есть в кластере окажутся сразу несколько документов для каждого изобретения. Тогда и в списке результатов поиска может быть один или несколько документов для каждого изобретения (или может не быть ни одного – в случае неудачного поиска). Если найден хотя бы один документ из патентного семейства изобретения, указанного в поле (56), то считаем, что найдено одно изобретение.

При вычислении оценки качества поиска используется количество найденных изобретений, а не просто количество найденных документов из кластера. Это и предлагается считать количеством найденных релевантных документов.

Таким образом, для автоматически сформированных наборов данных необходимо оценивать умение системы находить любые документы из патентных семейств, к которым могут быть отнесены документы, процитированные в поле (56). Вполне допустимо, что учёт даже всех патентных семейств для всех документов из поля (56) – как исчерпывающего набора релевантных документов – тоже может оказаться неверным, так как эксперт может процитировать в поле (56) только часть найденных релевантных документов или, в ряде случаев, остановить поиск, если посчитает, что нашёл достаточно. Следовательно, опираться на показатели, использующие TP, TN, FP и FN в любых вариациях нужно с большими оговорками. Возможно, более продуктивно будет оценивать систему в первую очередь с точки зрения того, способна ли она вообще извлекать релевантные документы – это как раз можно оценить, опираясь на содержимое поля (56) и патентные семейства указанных там документов.

Качество системы патентного поиска уровня техники предлагается оценивать по двум показателям [**29**].

Первый показатель оценивает способность системы извлекать релевантные запросу на поиск (то есть релевантные целевому документу) документы из предложенной коллекции и ставить их на первые K позиций ранжирования. Идея показателя схожа с известным показателем hit@K, применяемым для оценки систем, возвращающих несколько ответов. Hit@K – доля запросов в систему, для которых система вернула хотя бы один корректный ответ среди первых K [**22**, **30**].

Для одного запроса на поиск (целевого документа) показатель s принимает значение 1, если хотя бы один найденный релевантный документ попадает в top(K) списка найденных системой, в противном случае показатель принимает значение 0.

$$s = \begin{cases} 1, \text{ если хотя бы один релевантный документ есть в } top(K) \text{ выдачи;} \\ 0 \text{ в противном случае} \end{cases} \quad (1)$$

Для оценки качества работы системы на множестве запросов вычисляется арифметическое среднее показателя на документах запроса:

$$S@K = \frac{1}{N}\sum_{i=1}^{N}(s_i), \quad (2)$$

где N – количество целевых документов.

Показатель может быть интерпретирован как вероятность извлечения системой и представления в top(K) хотя бы одного релевантного документа из коллекции для каждого отдельно взятого выполненного поиска.

Второй показатель оценивает, насколько хорошо система уловила семантическое сходство целевого документа и документов из коллекции, в которой проводится поиск. При этом для одного запроса на поиск: в том случае, когда количество изобретений в патентном кластере больше или равно K, показатель h принимает значение 1, если все документы из top(K) списка найденных системой документов являются релевантными. Для количества изобретений в кластере меньше K, показатель принимает h значение 1, если как минимум по одному документу, характеризующему каждое изобретение в кластере, попали в top(K) списка найденных системой; во всех других случаях показатель принимает значение 0:

$$h = \begin{cases} 1, \text{ если количество изобретений в кластере} \leqslant K \\ \text{и все они попали в } top(K); \\ 1, \text{ если количество изобретений в кластере} > K \\ \text{и в } top(K) \text{ только релевантные документы;} \\ 0 \text{ во всех остальных документах} \end{cases} \quad (3)$$

Для оценки качества работы системы на множестве запросов вычисляется арифметическое среднее показателя на документах запроса, для первых K документов отзыва:

$$H@K = \frac{1}{N}\sum_{i=1}^{N}(h_i), \quad (4)$$

где N – количество целевых документов.

Показатель может быть интерпретирован как вероятность извлечения системой и ранжирования на top(K) позиции всех релевантных документов из коллекции для каждого отдельно взятого выполненного поиска.

3. Результаты и обсуждение

Для экспериментов с наборами данных, организованными по кластерному принципу, из коллекции Роспатента извлекались базовые документы по следующим правилам:

1. Патентные документы на русском языке: опубликованные Роспатентом документы за 1994–2020 гг. Общее количество документов в коллекции – 1 539 688;

2. Патентные документы на английском языке: общедоступные патентные документы США с датой публикации с 2000 до конца 2020 г., имеющие код вида документа A1 (заявка – Utility Patent Application published on or after January 2, 2001) или B2 (патент – Utility Patent Grant (with pre-grant publication) issued on or after January 2, 2001). В поисковой платформе Роспатента в массиве документов США по состоянию на 02.06.2023 всего документов в указанном диапазоне дат – 14 232 161, в том числе документов с кодами вида документа A1 или B2 всего 12 125 081 (из них патентов с кодом вида документа B2 – 4 033 783).

Для формирования наборов данных по кластерному принципу было разработано программное обеспечение генератора наборов данных [**31**]. Генератор позволяет строить наборы данных произвольной конфигурации, извлекая из предложенной ему коллекции базовые документы и формируя кластеры каждого базового документа.

Были построены наборы данных, содержащие кластеры русскоязычных документов для патентных документов на русском языке и кластеры англоязычных документов для патентных документов на английском языке. В обоих случаях исследовались возможность и характер расширения наборов данных документами из коллекции Роспатента – как в диапазоне указанных дат, так и за их пределами, при этом в кластеры включались не все документы и члены их патентных семейств, ссылки на которые содержались в поле (56) базового документа, а только документы на языке коллекции (русский и английский соответственно).

Для русскоязычного набора данных в среднем мощность кластера составила три и более документа, включая базовый документ. Англоязычный набор данных содержит кластеры, мощность которых в среднем превышает пятнадцать документов. Существенно более заметное расширение наборов данных объясняется в первую очередь обширностью коллекции на английском языке, а также более широким использованием английского языка в патентах. При этом в обоих случаях основное свойство кластера – объединить документы уровня техники для

базового документа – выполняется, а, значит, позволяет формировать наборы данных автоматизированным способом.

Характерной особенностью наборов данных, организованных по кластерному принципу, является повторение наиболее цитируемых патентов в разных кластерах. Выскажем гипотезу, что для поисковых систем, имеющих в своей основе тематическое моделирование [**32**, **33**], принципы дистрибутивной семантики и TF-IDF [**34**], такие наборы данных должны оказывать влияние на эффективность обучения системы.

Вторая гипотеза заключается в том, что кластеры документов уровня техники на всех языках, на которых они были опубликованы, должны в какой-то мере компенсировать недостаток в параллельных патентных текстах, особенно для русского языка, ощущающийся в сообществе. В настоящее время в Роспатенте готовятся наборы данных, состоящие из многоязычных кластеров патентных документов.

Запланированы эксперименты по оценке эффективности применения кластерных наборов данных и проверке двух указанных гипотез.

Для оценки качества поиска путём вычисления описанных выше показателей была подготовлена программная утилита [**31**]. Утилита работает совместно с генератором наборов данных и позволяет оценивать качество поиска по двум описанным выше критериям. В дополнение в утилите реализован расчёт двух наиболее широко используемых оценок качества поиска – точности поиска и полноты поиска, точнее их вариаций для первых K документов ранжирования: precision@K (P@K) и recall@K (R@K). Для множества запросов вычислялись усреднённая точность поиска (mean precision, MP@K) и усреднённая полнота поиска (mean recall MR@K).

Эксперименты показали более высокую информативность показателей P@K и R@K по сравнению с S и H (формулы 1 и 3) для одиночного запроса на поиск. Очевидно, это связано с бинарным характером S и H. Однако для множества поисков показатели S и H (формулы 2 и 4) дают больше информации о качестве поисковой системы, чем P и R, усреднённые по множеству поисков, независимо от того, какое «среднее» вычисляется. Показатель S напрямую оценивает долю «удачных» поисков на множестве запросов (S), считая «удачным» поиск с ненулевым извлечением релевантных документов. Долю исчерпывающе полных поисков определяет значение показателя H. В то же время MP и MR могут вводить пользователя в заблуждение, например, если систе-

ма демонстрирует высокие показатели для какой-то группы документов в множестве запросов, но полностью проваливается на других – в таком случае может оказаться, что сравнительно небольшое количество удачных результатов с высокой точностью/полнотой «замаскируют» слабую работу системы в целом.

Другой случай – выбор слишком большого K для какой-либо группы документов. Смысл параметра состоит в том, чтобы ограничить объём выборки документов, предлагаемых поисковой системой к рассмотрению эксперту. Для слишком маленького K есть риск, что важные документы уровня техники не попадут в выборку и не будут рассмотрены экспертом. При слишком большом K в выборку может попасть слишком много нерелевантных документов. Для патентного поиска автор и коллеги предложили принять K = 20 [**29**], считая при этом, что двадцати документов в большинстве случаев достаточно, чтобы оценить уровень техники. Для тех случаев, когда стопроцентная полнота (r@20 = 1) достигается на меньшем количестве документов, нерелевантные документы в выборке будет снижать p@20.

В этом смысле показатели S и H свободны от указанных недостатков. Сказанное демонстрируется в таблице на примере пятнадцати реальных кластеров документов ведомства США, отобранных случайным образом.

Продемонстрированы средние оценки (группа ячеек «Средние») и оценки показателей качества поиска для каждого документа (колонки «s@20», «h@20», «P@20» и «R@20»). При этом задачей поиска ставилось нахождение изобретений, запатентованных либо рассматриваемых ведомством США (патенты и заявки США), демонстрирующих уровень техники для патентов, вынесенных в колонку «Документы», то есть в результатах поиска релевантными цели поиска документами (колонка «Найдено релевантных документов US») считались только документы с кодом ведомства US. Поскольку, как уже было сказано выше, при поиске на уровень техники задачей является найти именно изобретения, то при оценивании нескольких документов, относящихся к одному изобретению, они учитывались как один документ/одно изобретение (колонка «Найдено изобретений»).

Пример различий показателей оценки качества поиска

Документ	S@20	H@20	P@20	R@20	Документов в кластере	Документов US	Изобретений	Найдено релевантных документов US	Найдено изобретений
US8446984B2	1	1	0,15	1	13	9	3	5	3
US7722993B2	1	1	0,1	1	36	3	2	3	2
US7929066B2	1	1	0,1	1	23	8	2	6	2
US8430280B2	1	0	0,05	0,2	40	6	5	2	1
US8683639B2	1	0	0,15	0,6	70	14	5	13	3
US7926563B2	1	0	0,2	0,8	36	10	5	7	4
US8147095B2	0	0	0	0	18	10	6	0	0
US7942420B2	0	0	0	0	23	8	7	0	0
US8174042B2	1	0	0,05	0,3333	262	55	3	1	1
US9032770B2	1	0	0,05	0,1429	69	13	7	3	1
US8735317B2	1	0	0,1	0,3333	78	10	6	2	2
US9025644B2	1	0	0,1	0,25	84	63	8	9	2
US9675911B2	0	0	0	0	27	5	1	0	0
US9599102B2	0	0	0	0	3	3	2	0	0
US8969517B2	0	0	0	0	15	3	2	0	0
Средние	0,67	0,20	0,07	0,3773					
	S@20	H@20	MP@20	MR@20					

Как видно из таблицы, при таких граничных условиях ни для одного документа не могло быть найдено 20 изобретений для демонстрации уровня техники – столько просто не было указано экспертами в поле (56) (колонка «Изобретений»). Следовательно, показатели P@20 и MP@20 дают заведомо искажённый результат. Показатель полноты для каждого отдельного поиска (R@20) информативнее, чем бинарные S@20 и H@20, но усреднённый показатель MR@20 мало что даёт для оценки эффективности поисков.

Показатели S@20 и H@20 легко интерпретируются как «Доля поисков, в которых найдено хотя бы одно релевантное изобретение» и «Доля поисков, в которых найдены все релевантные изобретения» соответственно.

4. Заключение

В работе описан новый подход к автоматизированному построению наборов данных патентных документов, основанный на формировании патентных кластеров – документов уровня техники для базового документа кластера. Введено понятие базового документа кластера и даны рекомендации по формированию кластеров.

Предложены методика и показатели оценки качества поиска, вычисляемые с учётом не только документов, отмеченных как релевантные экспертами патентного ведомства, но и учитывающая попадание документов кластера в ранжирование первых K документов отзыва поисковой системы.

Запланированы работы по созданию наборов данных на основе многоязычных кластеров патентных документов и эксперименты по применению таких наборов данных для обучения и тестирования автоматических поисковых систем для многоязычного патентного поиска.

Коллекции патентных документов, генератор наборов данных и утилита оценки качества поиска находятся в свободном доступе и предоставляются по запросу.

Список источников

1. **WIPO** Publication. WIPO Intellectual Property Handbook Second Edition, volume No. 489 (E). WIPO, 2004. ISBN 978-92-805-1291-5.

2. **Rubilar-Torrealba R., Chahuán-Jiménez K., de la Fuente-Mella H.** Analysis of the Growth in the Number of Patents Granted and Its Effect over the Level of Growth of the Countries: An Econometric Estimation of the Mixed Model Approach. Sustainability 2022, 14, 2384.

3. **OECD.** Patents and Innovation: Trends and Policy Challenges; OECD Organization for Economic Co-operation and Developmemt. Paris, France, 2004.

4. **Shalaby W., Zadrozny W.** Patent retrieval: A literature review. Knowl. Inf. Syst. 2019, 61, 631–660.

5. **Risch J., Krestel R.** Domain-specific word embeddings for patent classification. Data Technol. Appl. 2019, 53, 108–122.

6. **Pogiatzis A.** NLP: Contextualized Word Embeddings from BERT. 20 March 2019. URL: https://towardsdatascience.com/nlp-extract-contextualized-word-embeddings-from-bert-keras-tf-67ef29f60a7b.

7. **Humayun M. A., Yassin H., Shuja J., Alourani A., Abas P. E.** A transformer fine-tuning strategy for text dialect identification. Neural Comput. Appl. 2023, 35, 6115–6124.

8. **Roda G., Tait J., Piroi F., Zenz V.** CLEF-IP 2009: Retrieval Experiments in the Intellectual Property Domain. In Proceedings of the Workshop of the Cross-Language Evaluation Forum for European Languages. Corfu, Greece, 30 September – 2 October 2009. Volume 1175.

9. **Piroi F.** CLEF-IP 2010: Retrieval Experiments in the Intellectual Property Domain. In Proceedings of the CLEF 2010. Padua, Italy, 20–23 September 2010.

10. **Piroi F., Lupu M., Hanbury A., Zenz V.** CLEF-IP 2011: Retrieval in the intellectual property domain. In Proceedings of the CLEF 2011. Amsterdam, The Netherlands, 19–22 September 2011.

11. **Piroi F., Lupu M., Hanbury A., Magdy W., Sexton, A., Filippov I.** CLEF-IP 2012: Retrieval experiments in the intellectual property domain. In Proceedings of the CEURWorkshop. Melbourne, Australia, 10–12 December 2012; Proceedings 1178.

12. **Iwayama M., Fujii A., Kando N., Takano A.** Overview of patent retrieval task at NTCIR-3. In Proceedings of the CL-2003 Workshop on Patent Corpus Processing. Sapporo, Japan, 12 July 2003.

13. **Fujii A., Iwayama M., Kando N.** Overview of Patent Retrieval Task at NTCIR-4. In Proceedings of the NTCIR-4. Tokyo, Japan, 2–4 June 2004. URL: https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/PATENT/NTCIR4-OVPATENT-FujiiA.pdf.

14. **Fujii A., Iwayama M., Kando N.** Overview of Patent Retrieval Task at NTCIR-5. In Proceedings of the NTCIR-5. Tokyo, Japan, 6–9 December 2005. URL: https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/data/PATENT/NTCIR5-OV-PATENT-FujiiA-pp.pdf.

15. **Fujii A., Iwayama M., Kando N.** Overview of the Sixth NTCRWorkshop. In Proceedings of the NTCIR-6. Tokyo, Japan, 15–18 May 2007. URL: http://ntur.lib.ntu.edu.tw/retrieve/170726/26.pdf.

16. **Lupu M., Piroi F., Huang X., Zhu J., Tait J.** Overview of the TREC 2009 chemical IR track. In Proceedings of the TREC 2009. Gaithersburg, MD, USA, 17–20 November 2009.

17. **Lupu M., Tait J., Huang J., Zhu J.** TREC-CHEM 2010: Notebook Report. In Proceedings of the TREC 2010. Gaithersburg, MD, USA, 16–19 November 2010; NIST Special Publication, 500–294. URL: https://trec.nist.gov/pubs/trec19/papers/CHEM.OVERVIEW.pdf.

18. **Lupu M., Gurulingappa H., Filippov I., Zhao, J., Fluck J., Jacobs M., Huang J., Tait J.** Overview of the TREC 2011 Chemical IR track. In Proceedings of the TREC 2011. Gaithersburg, MD, USA, 15–18 November 2011.

19. **SIGIR'19**: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval July 2019. Pp. 1213–1216. https://doi.org/10.1145/3331184.3331346.

20. **Положение** о Государственном патентном фонде. URL: https://www.fips.ru/documents/npa-rf/prikazy-rospatenta/polozhenie-o-gosudarstvennom-patentnom-fonde.php#4.

21. **WIPO** Standard ST.96. Version 6.0 (approved by the Task Force, October 3, 2022), Standard ST.96 (Main Body). URL:

https://www.wipo.int/export/sites/www/standards/en/pdf/03-96-01.pdf, Annexes I to VII. https://www.wipo.int/standards/en/st96/v6-0/.

22. **Mohammad M. Rahman,·Chanchal K. Roy,·David Lo.** Automatic query reformulation for code search using crowdsourced knowledge. Empirical Software Engineering. URL: https: //doi.org/10.1007/s10664-018-9671-0 https://www.researchgate.net/publication/330514983.

23. **Prior** Art Candidates Search Task, Публикация международного исследовательского проекта IRF (Information Retrieval Facility). URL: https://www.ir-facility.org/prior-art-search1.

24. **The Effect** of Content-Equivalent Near-Duplicates on the Evaluation of Search Engines, Maik Fröbe, Jan Philipp Bittner, Martin Potthast, Matthias Hagen; 42nd European Conference on Information Retrieval (ECIR). Lissabon, April 14–17, 2020.

25. **Горбунов А. В., Генин Б. Л., Золкин Д. С.** Семантические кластеры патентных документов и генератор наборов данных для машинного обучения // Сборник трудов X Международной научно-практической конференции «Интеллектуальная инженерная экономика и Индустрия 5.0» (ИНПРОМ), 25–28 апреля 2024, Санкт-Петербург. В 2 т. Т. 2 / под ред. д-ра экон. наук Д. Г. Родионова, д-ра экон. наук А. В. Бабкина. Санкт-Петербург: Изд-во ПОЛИТЕХ-ПРЕСС, 2024. С. 457–461. ISBN 978-5-7422-8536-2.

26. **Kundu R.** F1 Score in Machine Learning: Intro & Calculation. Machine Learning. 16 December 2022. URL: https://www.v7labs.com/blog/f1-score-guide.

27. **Otten N. V.** Mean Average Precision Made Simple [Complete Guide]. 14 September 2023. URL: https://spotintelligence.com/2023/09/14/mean-average-precision/.

28. **Manning C. D., Raghavan P., Schütze H.** Introduction to Information Retrieval. Cambridge University Press: New York, NY, USA, 2008.

29. **Горбунов А. В., Генин Б. Л., Золкин Д. С.** Искусственный интеллект в работе патентных ведомств // Информационные ресурсы России. 2021. № 3. С. 18–23.

30. **Frome A., Corrado G. S., Shlens J. et al.** DeViSE: A Deep Visual-Semantic Embedding Model // Advances in Neural Information Processing Systems 26 (NIPS 2013), December 5–10, 2013, Harrah and Harveys. NV, USA. Curran Associates, Inc., 2013. P. 2121–2129.

31. **Горбунов А. В., Генин Б. Л., Золкин Д. С.** Задача выявления элементов семантического кластера патентных документов для поиска уровня техники // ISSN 0548-0019, НТИ. Сер. 1: Организация и методика информационной работы. 2023, № 8. С. 27–32.

32. **Kumaravel G., Sankaranarayanan S.** PQPS: Prior-Art Query-Based Patent Summarizer Using RBM and Bi-LSTM. Mob. Inf. Syst. 2021, 2021, 2497770.

33. **Zihayat M., Etwaroo R.** A non-factoid question answering system for prior art search. Expert Syst. Appl. 2021, 177, 114910.

34. **Pradeep.** Understanding TF-IDF in NLP: A Comprehensive Guide. Medium. 2023. URL: https://medium.com/@er.iit.pradeep09/understanding-tf-idf-in-nlp-a-comprehensive-guide-26707db0cec5.

Reference

1. **WIPO** Publication. WIPO Intellectual Property Handbook Second Edition, volume No. 489 (E). WIPO, 2004. ISBN 978-92-805-1291-5.

2. **Rubilar-Torrealba R., Chahuán-Jiménez K., de la Fuente-Mella H.** Analysis of the Growth in the Number of Patents Granted and Its Effect over the Level of Growth of the Countries: An Econometric Estimation of the Mixed Model Approach. Sustainability 2022, 14, 2384.

3. **OECD.** Patents and Innovation: Trends and Policy Challenges; OECD Organization for Economic Co-operation and Developmemt. Paris, France, 2004.

4. **Shalaby W., Zadrozny W.** Patent retrieval: A literature review. Knowl. Inf. Syst. 2019, 61, 631–660.

5. **Risch J., Krestel R.** Domain-specific word embeddings for patent classification. Data Technol. Appl. 2019, 53, 108–122.

6. **Pogiatzis A.** NLP: Contextualized Word Embeddings from BERT. 20 March 2019. URL: https://towardsdatascience.com/nlp-extract-contextualized-word-embeddings-from-bert-keras-tf-67ef29f60a7b.

7. **Humayun M. A., Yassin H., Shuja J., Alourani A., Abas P. E.** A transformer fine-tuning strategy for text dialect identification. Neural Comput. Appl. 2023, 35, 6115–6124.

8. **Roda G., Tait J., Piroi F., Zenz V.** CLEF-IP 2009: Retrieval Experiments in the Intellectual Property Domain. In Proceedings of the Workshop of the Cross-Language Evaluation Forum for European Languages. Corfu, Greece, 30 September – 2 October 2009. Volume 1175.

9. **Piroi F.** CLEF-IP 2010: Retrieval Experiments in the Intellectual Property Domain. In Proceedings of the CLEF 2010. Padua, Italy, 20–23 September 2010.

10. **Piroi F., Lupu M., Hanbury A., Zenz V.** CLEF-IP 2011: Retrieval in the intellectual property domain. In Proceedings of the CLEF 2011. Amsterdam, The Netherlands, 19–22 September 2011.

11. **Piroi F., Lupu M., Hanbury A., Magdy W., Sexton, A., Filippov I.** CLEF-IP 2012: Retrieval experiments in the intellectual property domain. In Proceedings of the CEURWorkshop. Melbourne, Australia, 10–12 December 2012; Proceedings 1178.

12. **Iwayama M., Fujii A., Kando N., Takano A.** Overview of patent retrieval task at NTCIR-3. In Proceedings of the CL-2003 Workshop on Patent Corpus Processing. Sapporo, Japan, 12 July 2003.

13. **Fujii A., Iwayama M., Kando N.** Overview of Patent Retrieval Task at NTCIR-4. In Proceedings of the NTCIR-4. Tokyo, Japan, 2–4 June 2004. URL: https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/PATENT/NTCIR4-OVPATENT-FujiiA.pdf.

14. **Fujii A., Iwayama M., Kando N.** Overview of Patent Retrieval Task at NTCIR-5. In Proceedings of the NTCIR-5. Tokyo, Japan, 6–9 December 2005. URL: https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/data/PATENT/NTCIR5-OV-PATENT-FujiiA-pp.pdf.

15. **Fujii A., Iwayama M., Kando N.** Overview of the Sixth NTCRWorkshop. In Proceedings of the NTCIR-6. Tokyo, Japan, 15–18 May 2007. URL: http://ntur.lib.ntu.edu.tw/retrieve/170726/26.pdf.

16. **Lupu M., Piroi F., Huang X., Zhu J., Tait J.** Overview of the TREC 2009 chemical IR track. In Proceedings of the TREC 2009. Gaithersburg, MD, USA, 17–20 November 2009.

17. **Lupu M., Tait J., Huang J., Zhu J.** TREC-CHEM 2010: Notebook Report. In Proceedings of the TREC 2010. Gaithersburg, MD, USA, 16–19 November 2010; NIST Special Publication, 500–294. URL: https://trec.nist.gov/pubs/trec19/papers/CHEM.OVERVIEW.pdf.

18. **Lupu M., Gurulingappa H., Filippov I., Zhao, J., Fluck J., Jacobs M., Huang J., Tait J.** Overview of the TREC 2011 Chemical IR track. In Proceedings of the TREC 2011. Gaithersburg, MD, USA, 15–18 November 2011.

19. **SIGIR'19**: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval July 2019. Pp. 1213–1216. https://doi.org/10.1145/3331184.3331346.

20. **Polozhenie** o Gosudarstvennom patentnom fonde. URL: https://www.fips.ru/documents/npa-rf/prikazy-rospatenta/polozhenie-o-gosudarstvennom-patentnom-fonde.php#4.

21. **WIPO** Standard ST.96. Version 6.0 (approved by the Task Force, October 3, 2022), Standard ST.96 (Main Body). URL: https://www.wipo.int/export/sites/www/standards/en/pdf/03-96-01.pdf, Annexes I to VII. https://www.wipo.int/standards/en/st96/v6-0/.

22. **Mohammad M. Rahman,·Chanchal K. Roy,·David Lo.** Automatic query reformulation for code search using crowdsourced knowledge. Empirical Software Engineering.

URL: https: //doi.org/10.1007/s10664-018-9671-0 https://www.researchgate.net/publication/330514983.

23. **Prior** Art Candidates Search Task, Публикация международного исследовательского проекта IRF (Information Retrieval Facility). URL: https://www.ir-facility.org/prior-art-search1.

24. **The Effect** of Content-Equivalent Near-Duplicates on the Evaluation of Search Engines, Maik Fröbe, Jan Philipp Bittner, Martin Potthast, Matthias Hagen; 42nd European Conference on Information Retrieval (ECIR). Lissabon, April 14–17, 2020.

25. **Gorbunov A. V., Genin B. L., Zolkin D. S.** Semanticheskie clastery` patentny`kh dokumentov i generator naborov danny`kh dlia mashinnogo obucheniia // Sbornik trudov X Mezhdunarodnoi` nauchno-prakticheskoi` konferentcii «Intellektual`naia inzhenernaia e`konomika i Industriia 5.0» (INPROM), 25–28 aprelia 2024, Sankt-Peterburg. V 2 t. T. 2 / pod red. d-ra e`kon. nauk D. G. Rodionova, d-ra e`kon. nauk A. V. Babkina. Sankt-Peterburg: Izd-vo POLITEKH-PRESS, 2024. S. 457–461. ISBN 978-5-7422-8536-2.

26. **Kundu R.** F1 Score in Machine Learning: Intro & Calculation. Machine Learning. 16 December 2022. URL: https://www.v7labs.com/blog/f1-score-guide.

27. **Otten N. V.** Mean Average Precision Made Simple [Complete Guide]. 14 September 2023. URL: https://spotintelligence.com/2023/09/14/mean-average-precision/.

28. **Manning C. D., Raghavan P., Schütze H.** Introduction to Information Retrieval. Cambridge University Press: New York, NY, USA, 2008.

29. **Gorbunov A. V., Genin B. L., Zolkin D. S.** Iskusstvenny`i` intellekt v rabote patentny`kh vedomstv // Informatcionny`e resursy` Rossii. 2021. № 3. S. 18–23.

30. **Frome A., Corrado G. S., Shlens J. et al.** DeViSE: A Deep Visual-Semantic Embedding Model // Advances in Neural Information Processing Systems 26 (NIPS 2013), December 5–10, 2013, Harrah and Harveys. NV, USA. Curran Associates, Inc., 2013. P. 2121–2129.

31. **Gorbunov A. V., Genin B. L., Zolkin D. S.** Zadacha vy`iavleniia e`lementov semanticheskogo clastera patentny`kh dokumentov dlia poiska urovnia tekhniki // ISSN 0548-0019, NTI. Ser. 1: Organizatciia i metodika informatcionnoi` raboty`. 2023, № 8. S. 27–32.

32. **Kumaravel G., Sankaranarayanan S.** PQPS: Prior-Art Query-Based Patent Summarizer Using RBM and Bi-LSTM. Mob. Inf. Syst. 2021, 2021, 2497770.

33. **Zihayat M., Etwaroo R.** A non-factoid question answering system for prior art search. Expert Syst. Appl. 2021, 177, 114910.

34. **Pradeep.** Understanding TF-IDF in NLP: A Comprehensive Guide. Medium. 2023. URL: https://medium.com/@er.iit.pradeep09/understanding-tf-idf-in-nlp-a-comprehensive-guide-26707db0cec5.

Информация об авторе / Author

Горбунов Александр Владимирович – начальник Центра развития научного направления «Искусственный интеллект» Федерального института промышленной собственности, Москва, Российская Федерация

gorbunov@rupto.ru

Alexander V. Gorbunov – Head, National Center for Artificial Intelligence, Federal Institute of Industrial Property, Moscow, Russian Federation

gorbunov@rupto.ru